THE UNIVERSITY of EDINBURGH

CHARLES UNIVERSITY

Deep Architectures for Neural Machine Translation

Antonio Valerio Miceli Barone[†] Jindřich Helcl[⋆] Rico Sennrich[†]
Barry Haddow[†] Alexandra Birch[†]

[†]School of Informatics, University of Edinburgh
[⋆]Faculty of Mathematics and Physics, Charles University

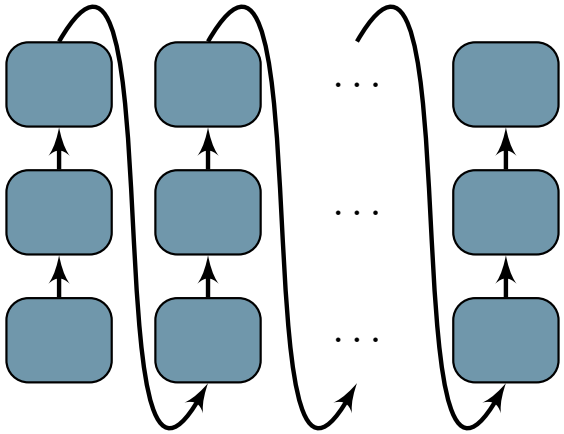
September 8, 2017

Deep architectures

- What is the depth of a recurrent neural network?

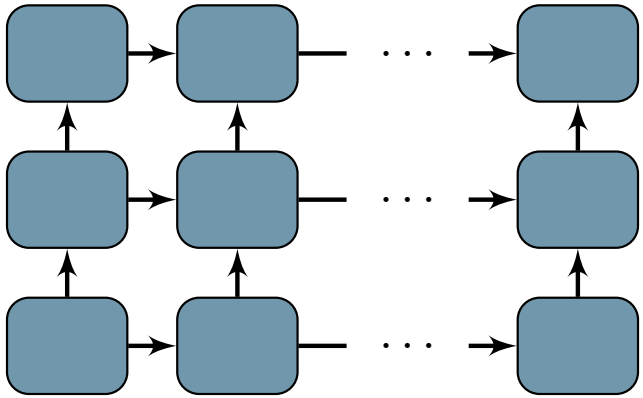
Deep architectures

- What is the depth of a recurrent neural network? [Pascanu et al., 2014]



Transition depth

- Used for LM [Zilly et al., 2016]

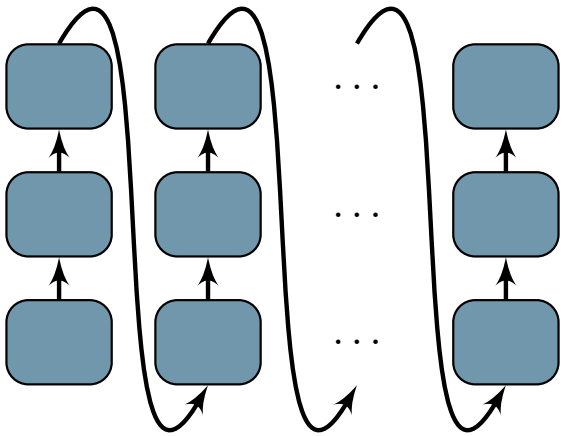


Stacked depth

- Baidu [Zhou et al., 2016]
- Google [Wu et al., 2016]

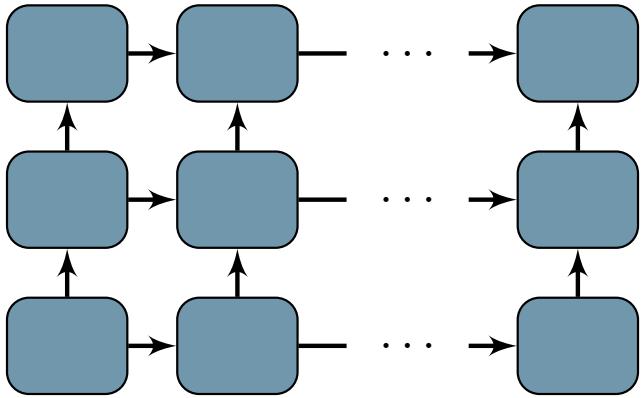
Deep architectures

- What is the depth of a recurrent neural network? [Pascanu et al., 2014]



Transition depth

- Used for LM [Zilly et al., 2016]

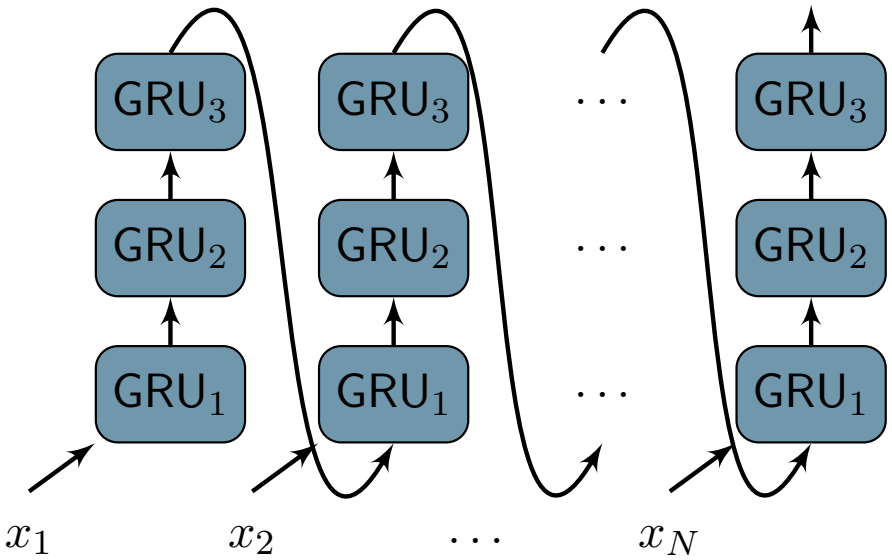


Stacked depth

- Baidu [Zhou et al., 2016]
- Google [Wu et al., 2016]

- This work
  - Provide a systematic comparison of deep architectures for NMT
  - Investigate the effects of different kinds of depth
  - Propose a combined "BiDeep" architecture

Deep transition encoder



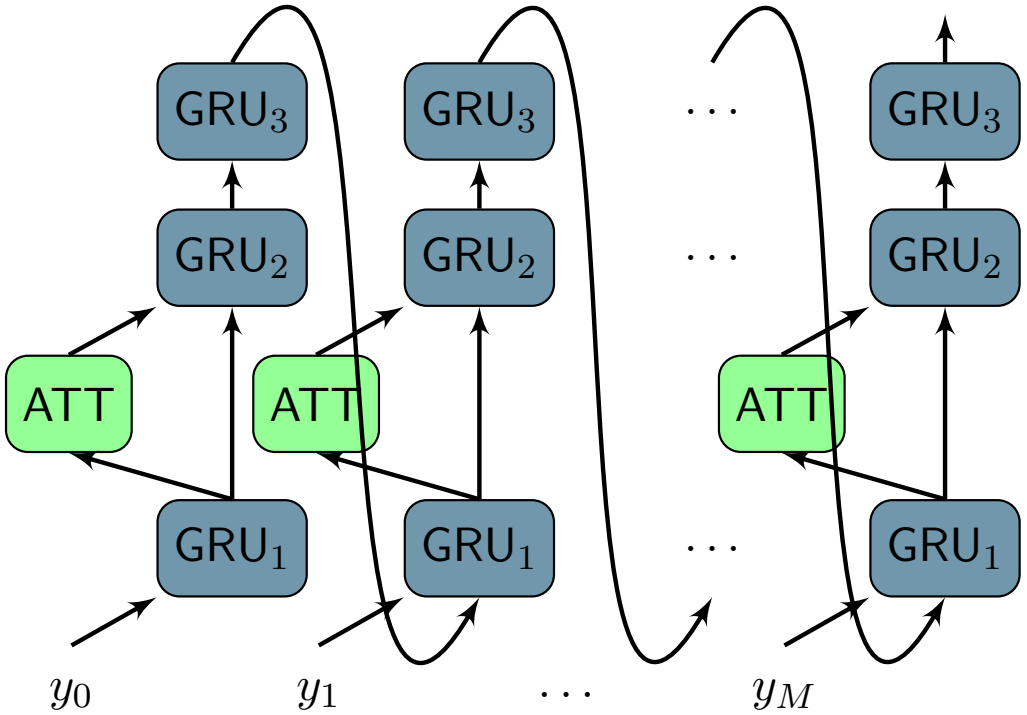
$$\overrightarrow{h}_{i,1} = \text{GRU}_1\left(x_i, \overrightarrow{h}_{i-1,L_s}\right)$$

$$\overrightarrow{h}_{i,k} = \text{GRU}_k\left(0, \overrightarrow{h}_{i,k-1}\right)$$

$$\text{for } 1 < k \leq L_s$$

- Bidirectional encoder
- Compute the next state using a deep feed-forward network made of multiple GRU transition blocks
- GRU blocks not individually recurrent, recurrence at time-step level

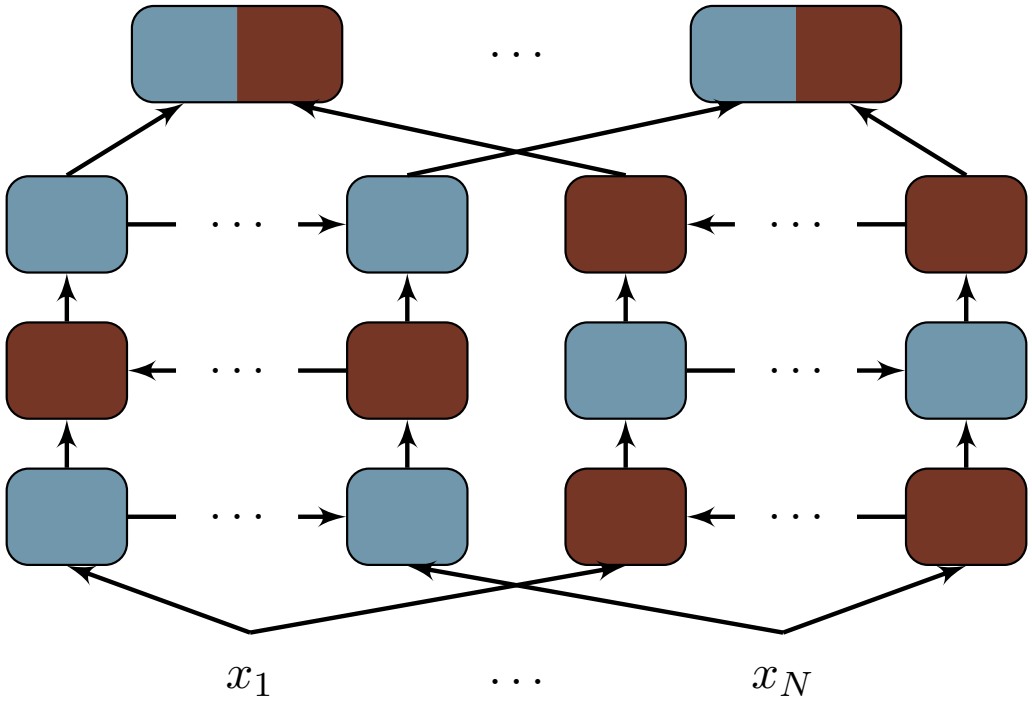
Deep transition decoder



$$s_{j,1} = \mathsf{GRU}_1\left(y_{j-1}, s_{j-1,L_t}\right)$$
$$s_{j,2} = \mathsf{GRU}_2\left(\mathsf{ATT}(C, s_{j,1}), s_{j,1}\right)$$
$$s_{j,k} = \mathsf{GRU}_k\left(0, s_{j,k-1}\right)$$
$$\text{for } 2 < k \leq L_t$$

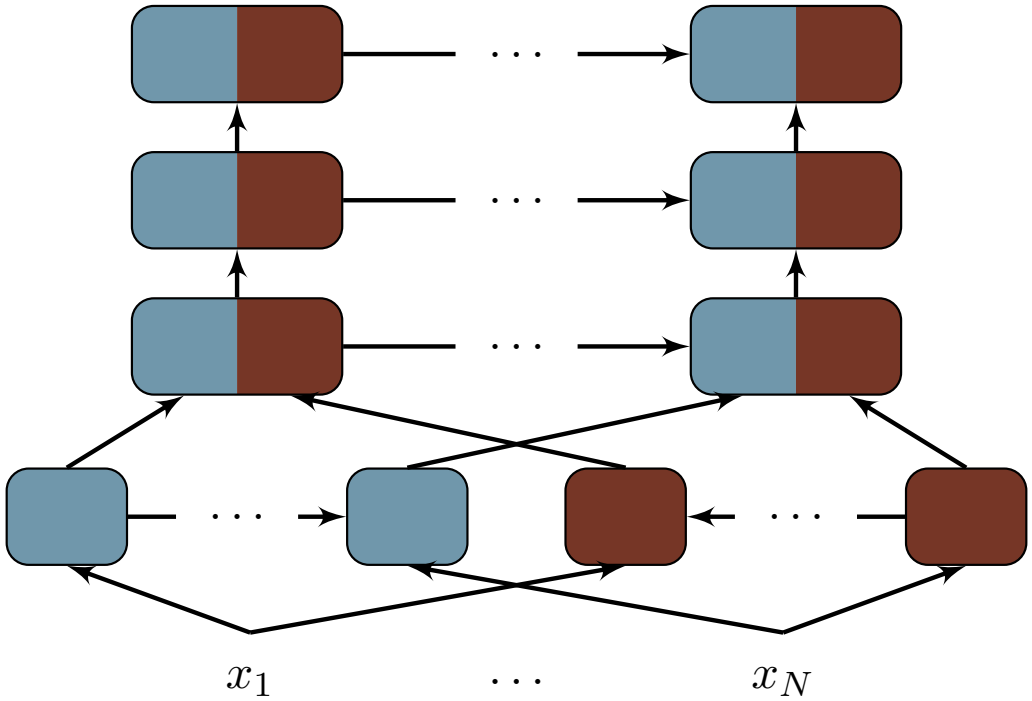
- Attention mechanism between 1st and 2nd layers (Nematus)
- Minimum transition depth is 2 even in the baseline

Alternating stacked encoder



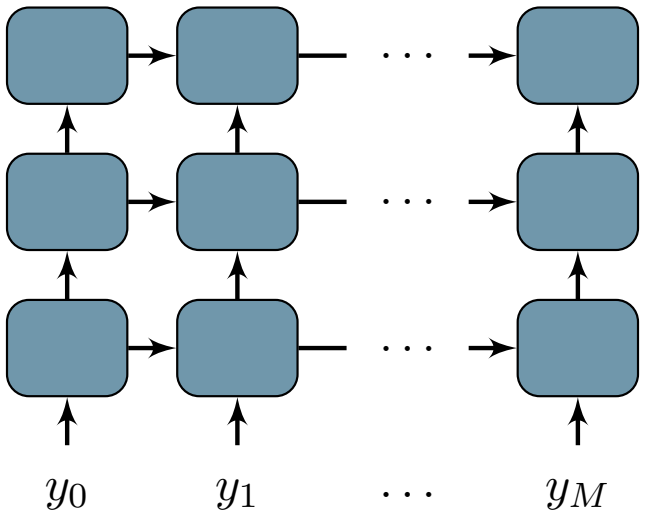
- Baidu [Zhou et al., 2016]
- Multiple levels of individually recurrent GRU cells
- Residual connections between levels
- Bidirectional: two columns with opposing scanning directions
- Each level inverts scanning direction of the previous one

Biunidirectional stacked encoder



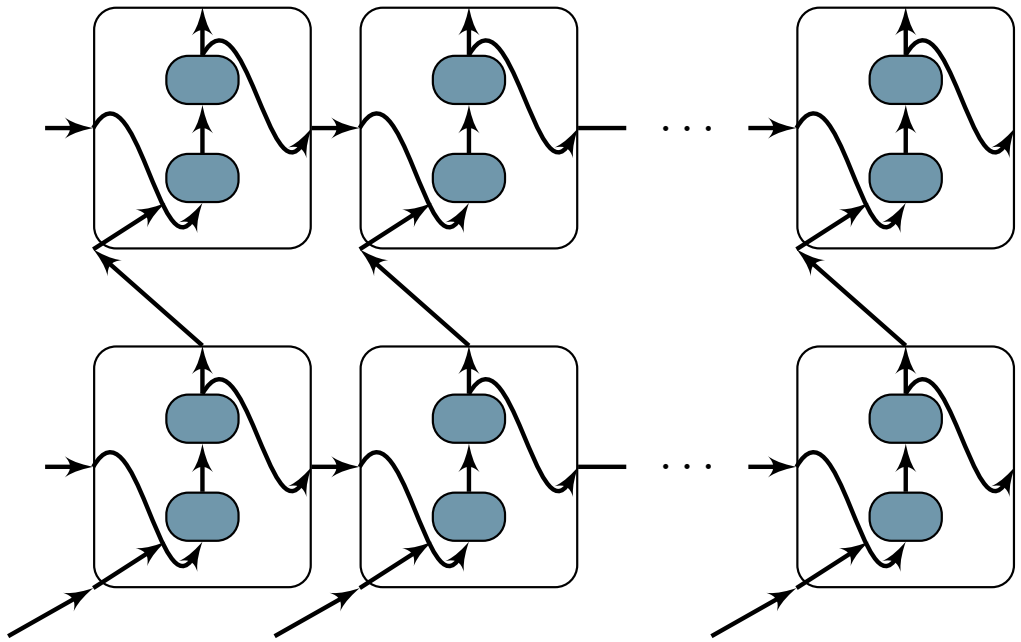
- Google [Wu et al., 2016]
- Multiple levels of individually recurrent GRU cells
- Residual connections between levels
- First levels are bidirectional, then states are merged
- Higher levels are unidirectional left-to-right

Stacked decoder



- Multiple levels of individually recurrent GRU cells
- Residual connections between levels
- Different variations depending on how attention is used in the higher layers
- (details in the paper)

BiDeep Architectures

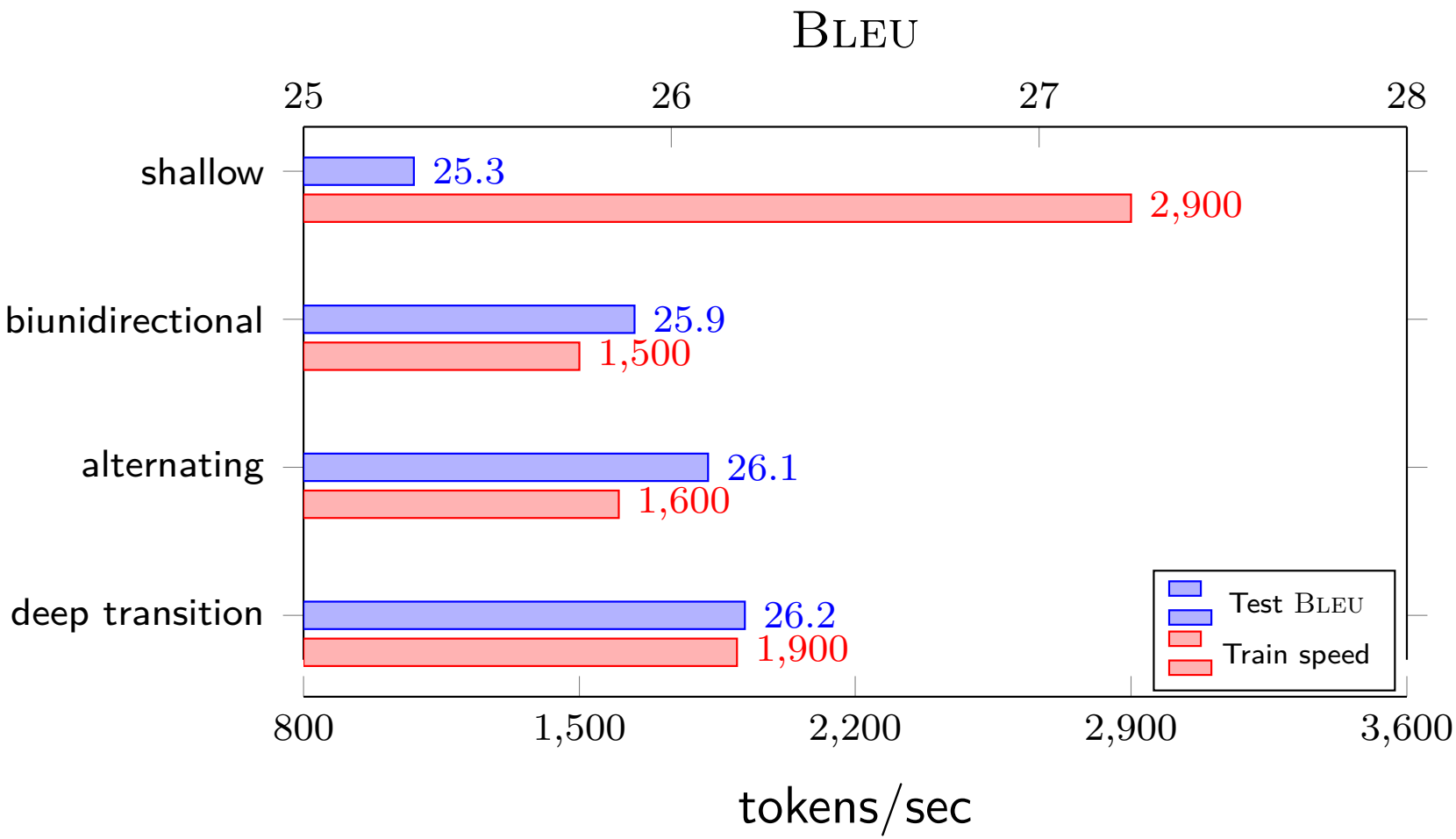


- Our proposal
- Combine the two kinds of depth
- Stacked levels of recurrent cells, each with multiple layers of transition depth

Experimental setup

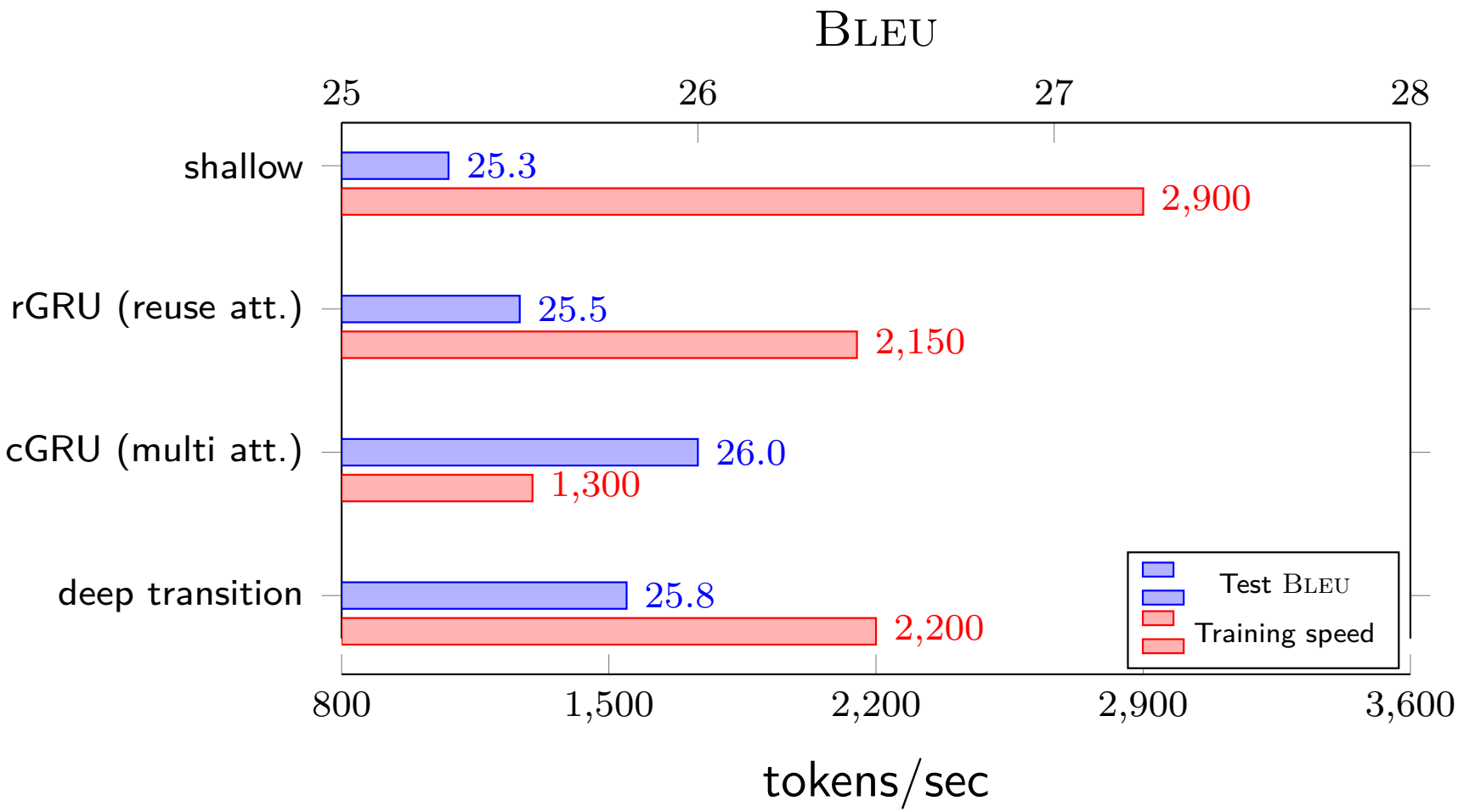
- Data
  - Training: WMT-2017 English-to-German
  - Validation: newstest 2013
  - Test: newstest 2014+2015+2016 (we report averages)
- System: Nematus [Sennrich et al., 2017b]
  - GRU sequence-to-sequence with attention [Bahdanau et al., 2015]
  - Layer normalization [Ba et al., 2016]
- Training on single Titan X (Pascal) GPU

Results: Deep Encoders



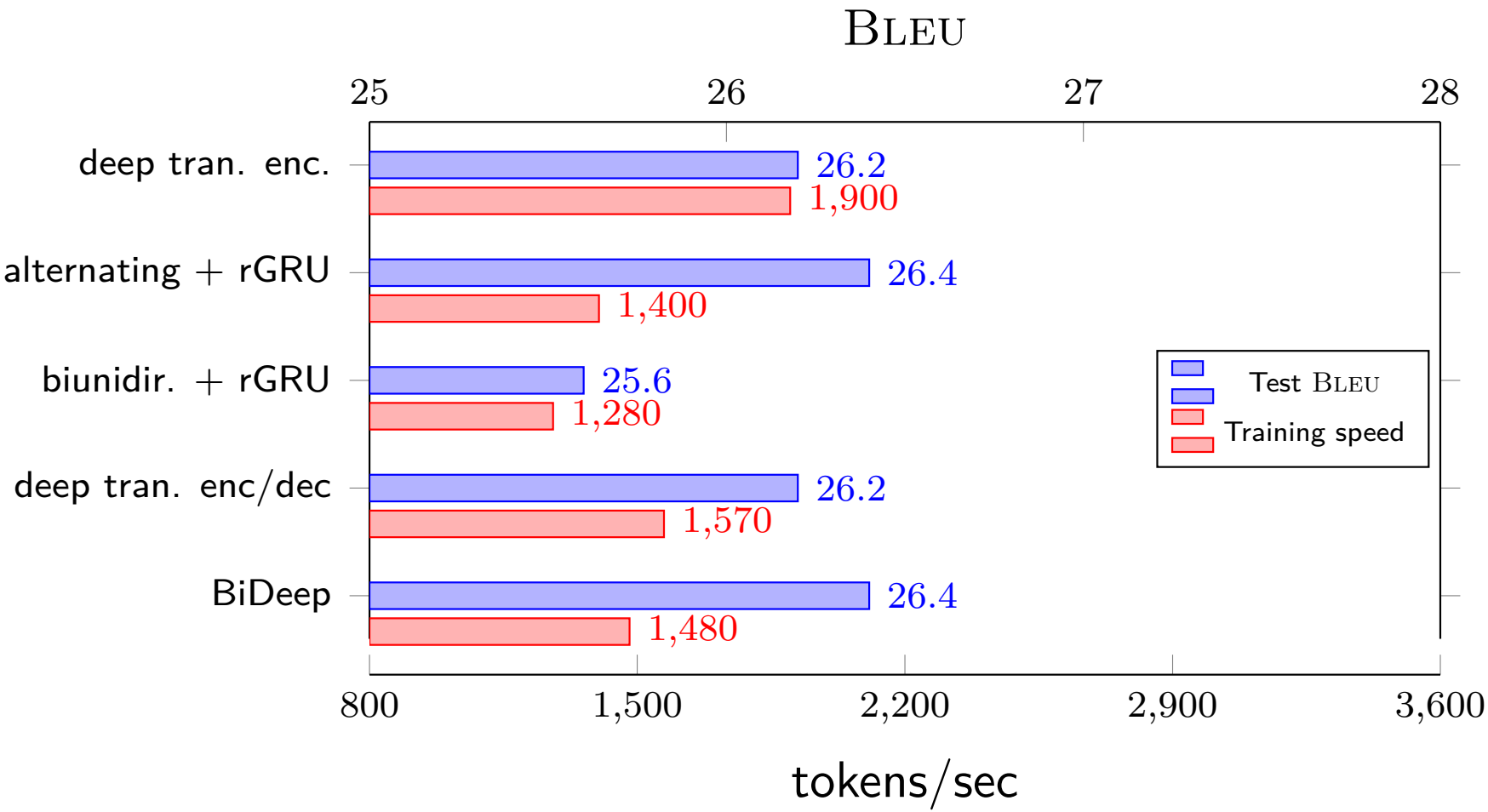
- Depth-4 encoders improve translation quality
- Deep transition fastest and most accurate

Results: Deep Decoders



- Depth-4 decoders improve translation quality
- Deep transition fastest and second most accurate

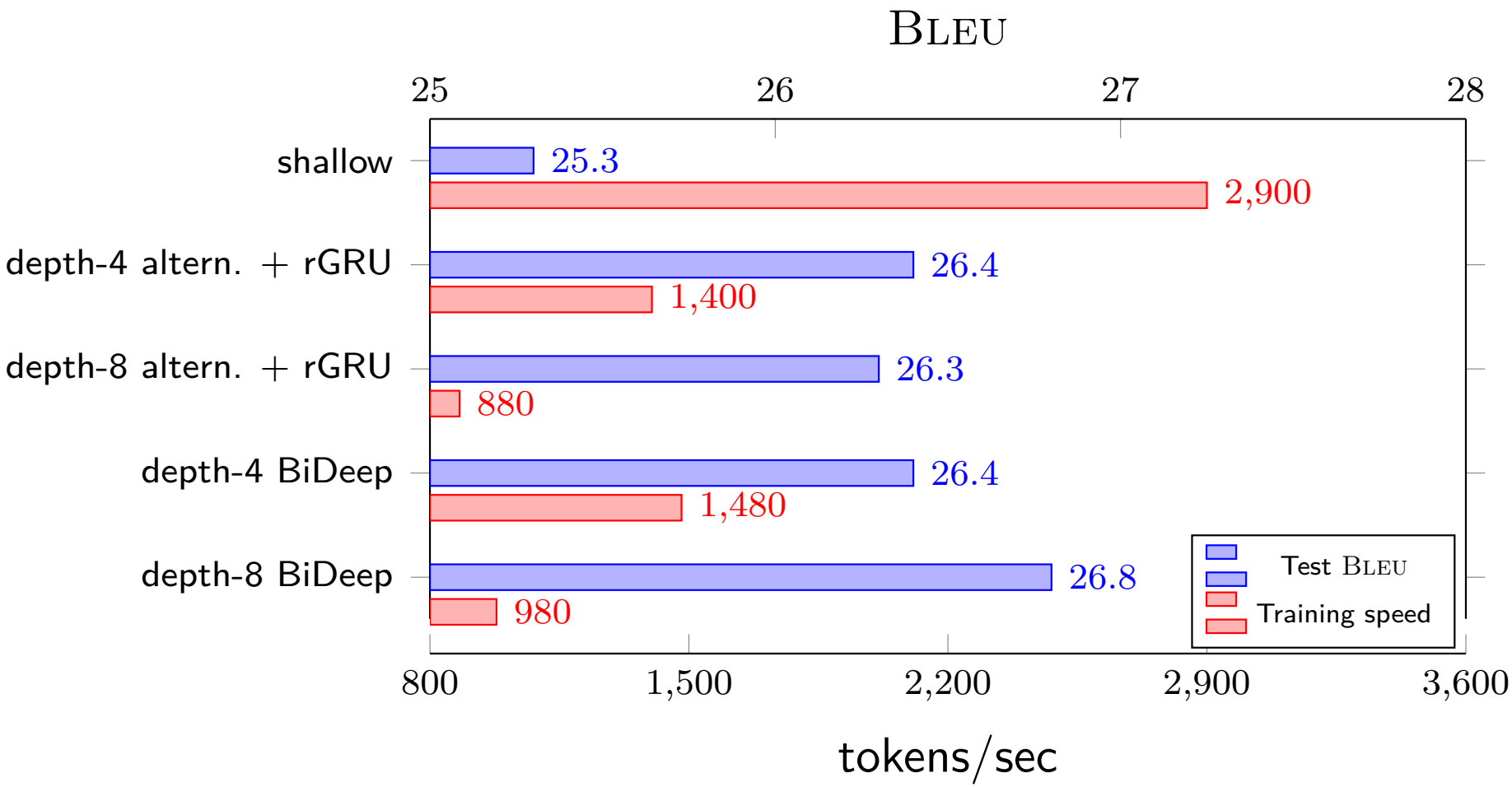
Results: Deep Encoders and Decoders



- Depth-4 on both encoder and decoder yields small improvement
- Biunidirectional+rGRU ≈[Wu et al., 2016] performs the worst
- Alternating+rGRU ≈[Zhou et al., 2016] and BiDeep are tied at this depth


STACK
MORE
LAYERS
LAYERS
LAYERS

Results: Deep Encoders and Decoders (depth 8)



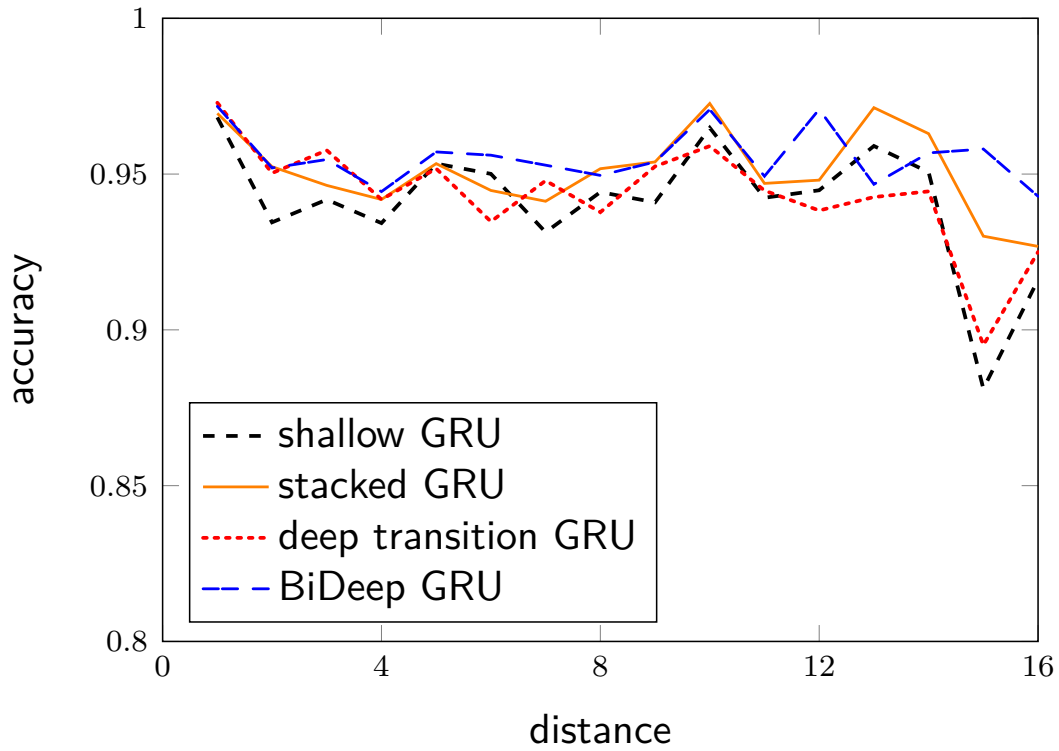
- Stacked-only plateaus
- BiDeep keeps improving

Error analysis

- Deep transition decoders have a longer information path
- In principle, might cause fading memory and vanishing gradients
- Does this affect long-distance dependencies?

Error analysis

- Deep transition decoders have a longer information path
- In principle, might cause fading memory and vanishing gradients
- Does this affect long-distance dependencies?
- Lingeval97 subject-verb agreement [Sennrich, 2017]
- Contrastive evaluation



Conclusions

- Findings
  - Depth improves translation BLEU especially in the encoder
  - Alternating stacked encoders outperform Biunidirectional
  - Deep transition encoders performs better or equal
  - BiDeep architectures perform the best
  - We validated these findings on the WMT-17 news translation task

Conclusions

- Findings
  - Depth improves translation BLEU especially in the encoder
  - Alternating stacked encoders outperform Biunidirectional
  - Deep transition encoders performs better or equal
  - BiDeep architectures perform the best
  - We validated these findings on the WMT-17 news translation task
- Recommendations
  - Use deep transition for speed and model size
  - Use BiDeep for maximum quality

Conclusions

- Code in the main **Nematus** repository
- Scripts and paper: https://git.io/v5W2Q

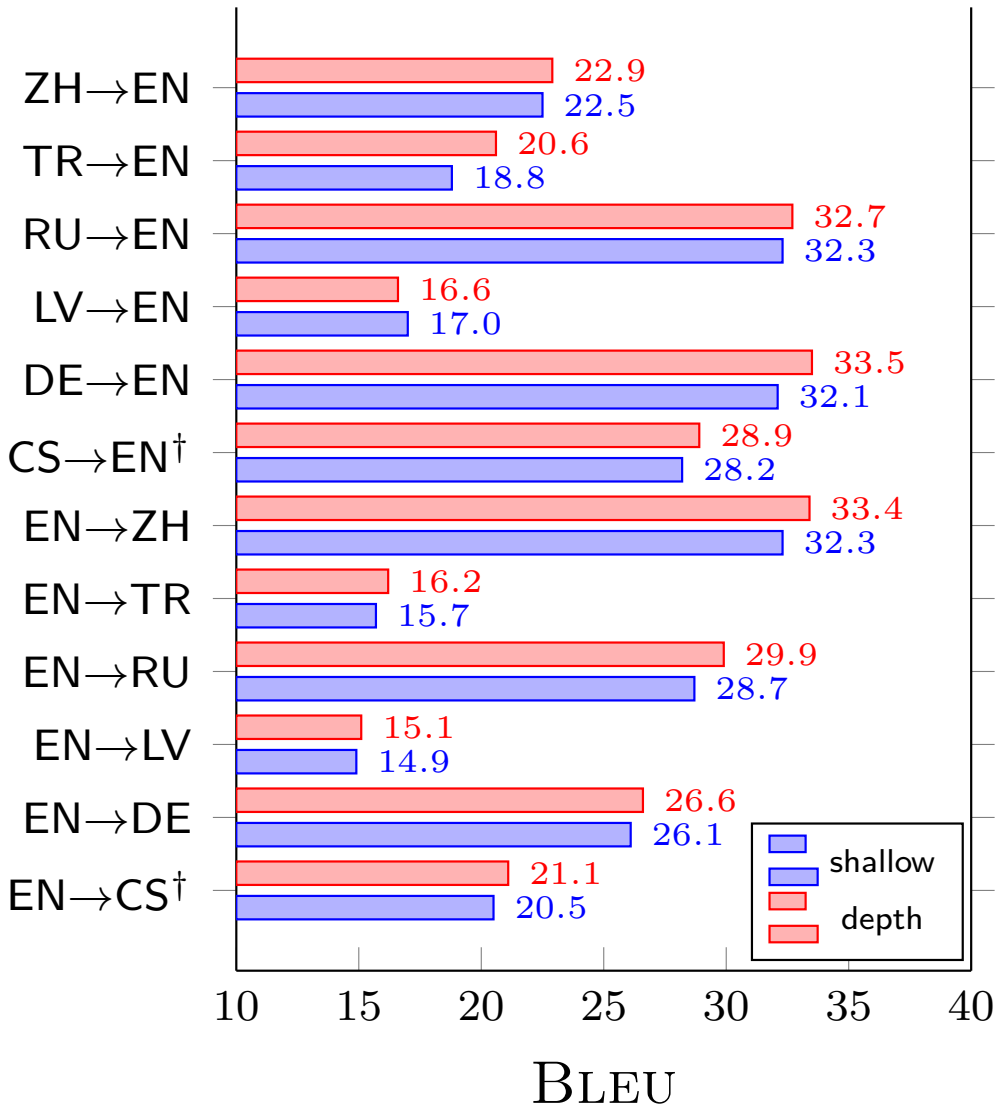
Thanks for your attention

Conclusions

- Code in the main **Nematus** repository
- Scripts and paper: https://git.io/v5W2Q

Thanks for your attention
Questions?

Results: WMT 2017 news translation task



- Transition depth: 8 + 4
- † Czech: stacked
- Improvement on all language pairs except Latvian↔English